IIG-HPC Cluster User Manual

256 Cores Intel HPC Cluster

Prepared by

Amar Kakad, Bharati Kakad and Mahendra Doiphode

Note: This is a preliminary document prepared for the IIG HPC system user. All IIG HPC users are welcome to contribute to this document. Kindly contact computer section for any corrections or suggestions.

Index

1. Specification

This document gives basic instructions to the new users of the cluster about logging in, compilation, running jobs, monitoring and cleaning jobs which did not exit gracefully.

The IIG clusters can be used in multiple ways. The clusters are named iighpc (Intel(R) Xeon based 64 bit). iighpc uses intel mpi over Infiniband interconnect. iighpc cluster consists of 8 compute nodes & 1 Head node. Supports both the openmp and MPI modes of programming.

Hardware Specification:-

Head Node: Fujitsu RX300 S8 – 1 node

CPU	2 numbers Intel XEON processor, 2 x 6 total 12 cores
RAM	8 x 8 GB total 64GB
Connectivity	Infiniband & Ethernet

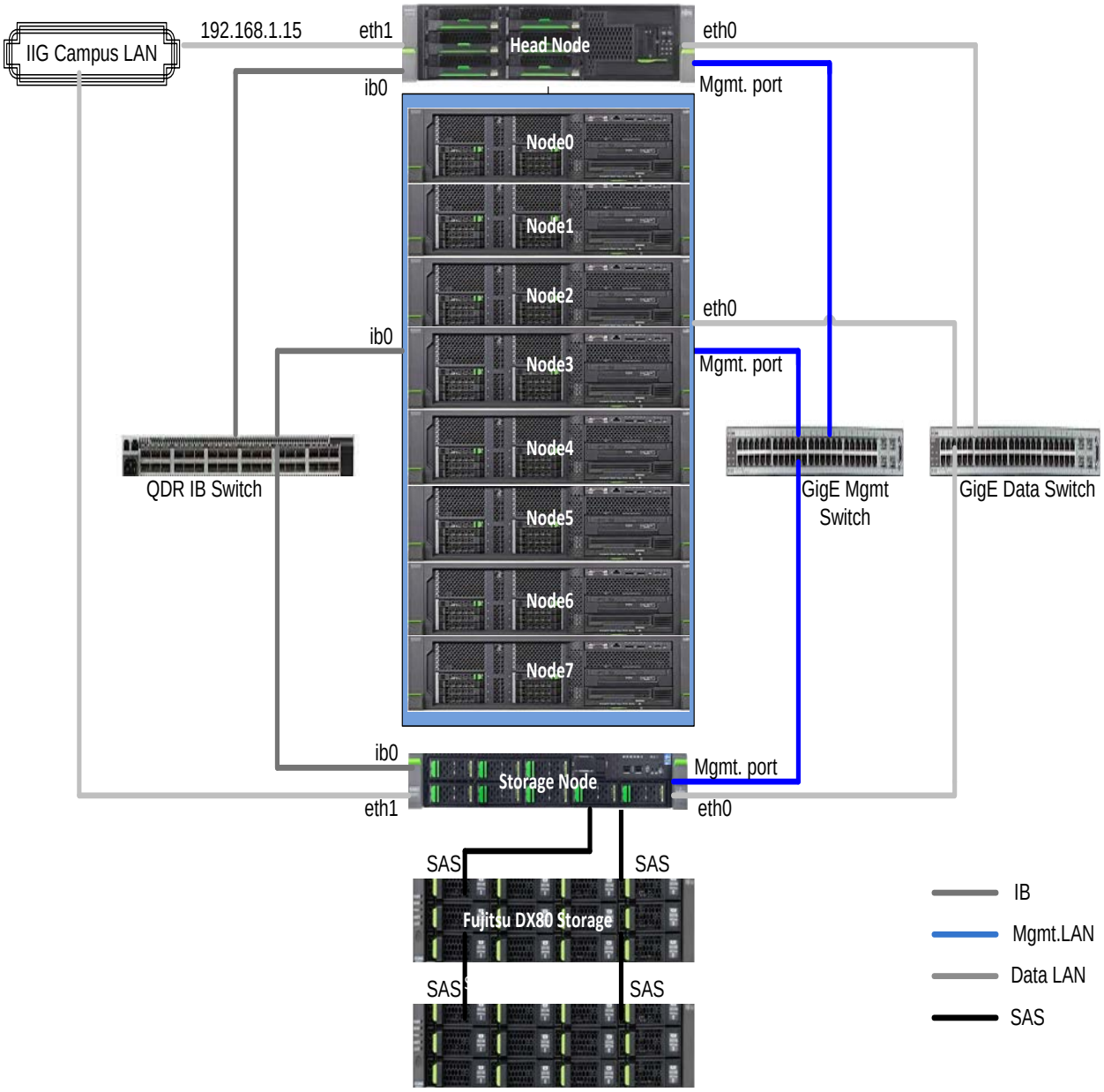
Compute Nodes: Fujitsu RX500 S7 - Total 8 nodes

CPU	4 numbers Intel XEON processor, 4 x 8 total 32 cores /node
RAM	16 x 16 GB total 256 GB/node
Connectivity	Infiniband & Ethernet

Software Specification:-

Operating System	RHEL 6.4, Kernel 2.6.32-358.el6.x86_64
Cluster Management Toolkit	GANANA (Locuz proprietary)
Scheduler	Univa Grid Engine (uge-8.1)
Compilers	GNU, Intel Cluster Studio XE 2013 SP1
Libraries	GNU, Intel Cluster Studio XE 2013 SP1
MPI	Intel MPI, Openmpi
OFED	Intel OFED

HPC Cluster Architecture



2. User Login

There are two ways for login to the cluster.

1. From windows machine

1) Using putty command line interface software

   Putty download link: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html

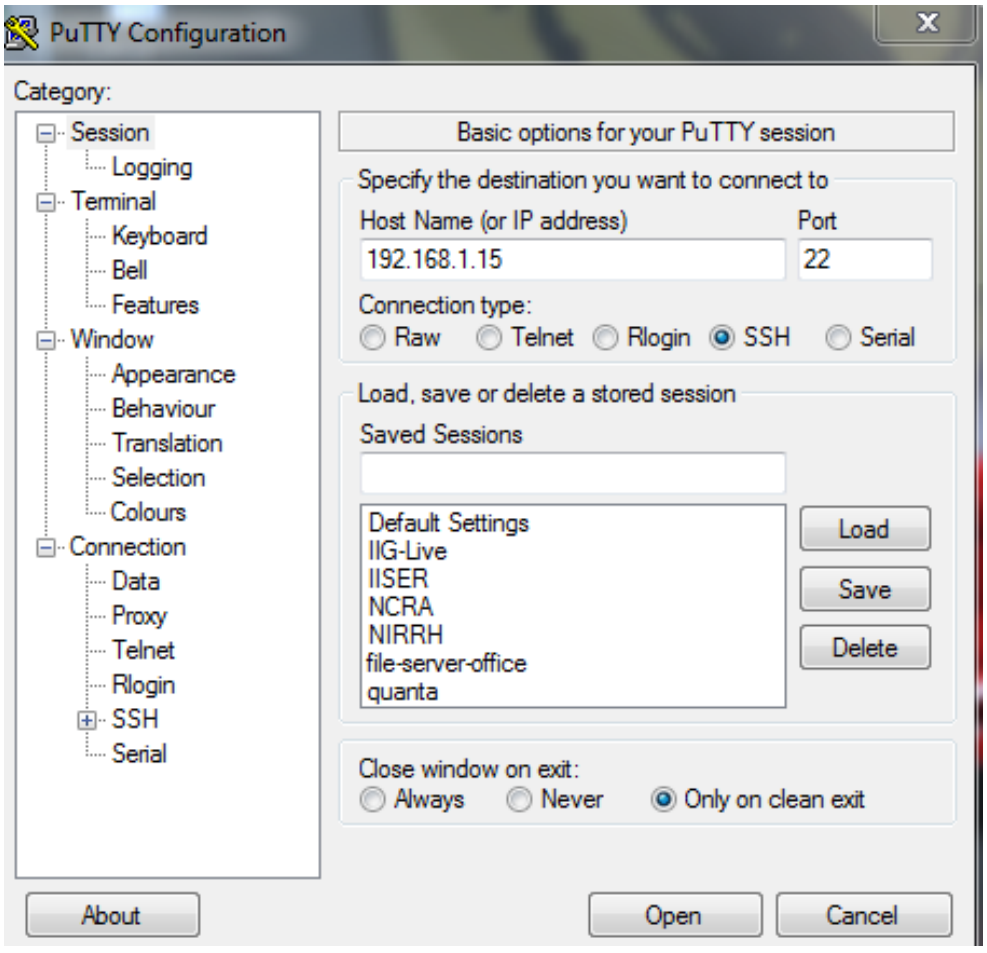
2) Using VNC Viewer

   VNC Viewer download link: http://www.realvnc.com/download/viewer

How to login using putty software

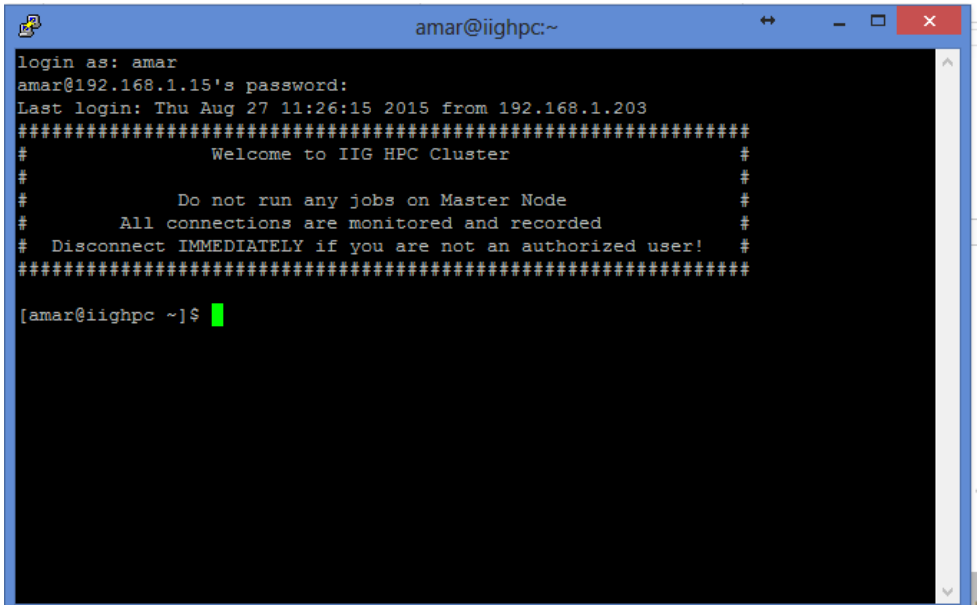
Open putty software & type: Head Node IP address with **port no. 22**

Note: - Public IP address of Head Node: **192.168.1.15**



Click on Open

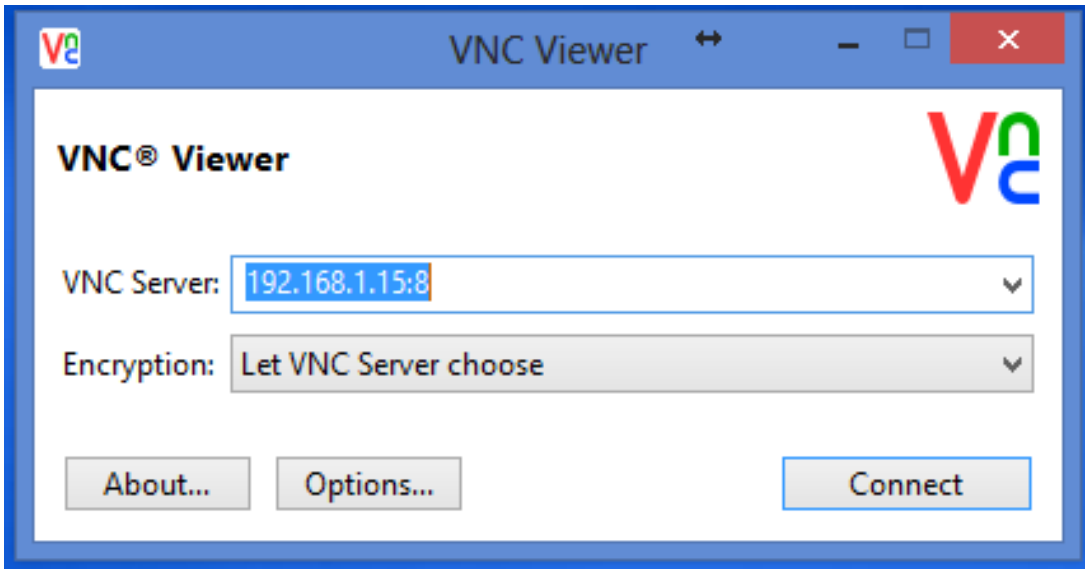
Next window will ask you about username & password. Entering valid login and password will get HPC access.

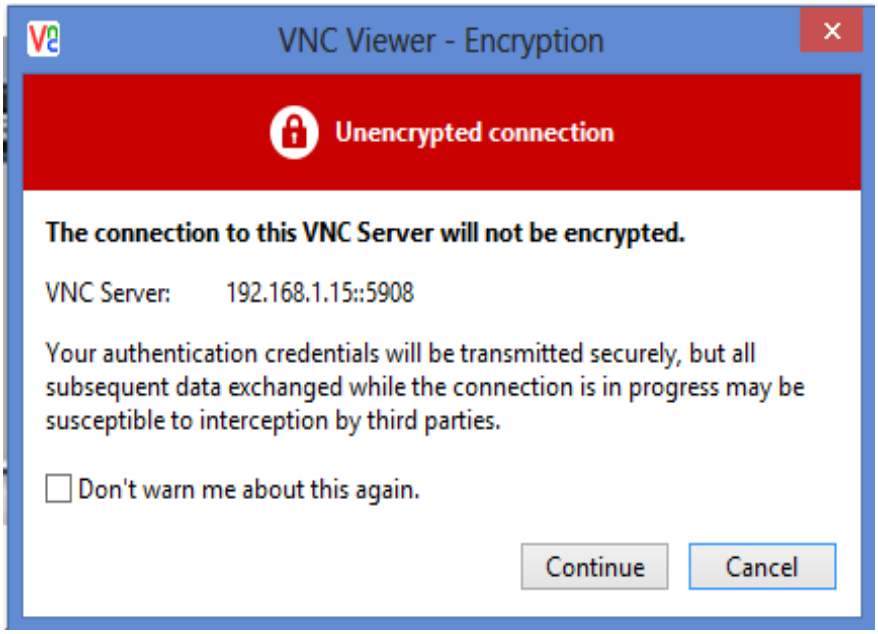


How to login using VNC Viewer software

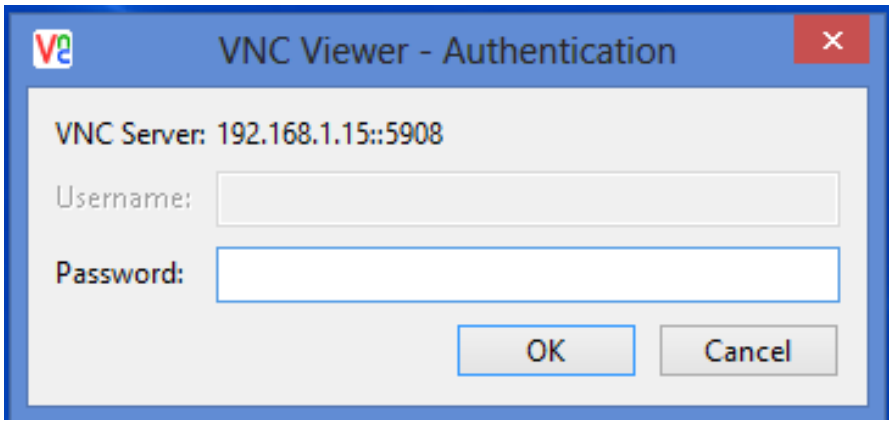
VNC Viewer software must be installed, which is freely available on the website https://www.realvnc.com/download/viewer/

From your desktop or laptop, Open VNC viewer software specify IP address of Head node along with port number which is unique for each user. This port number will be provided by the system administrator along with login ID and password.

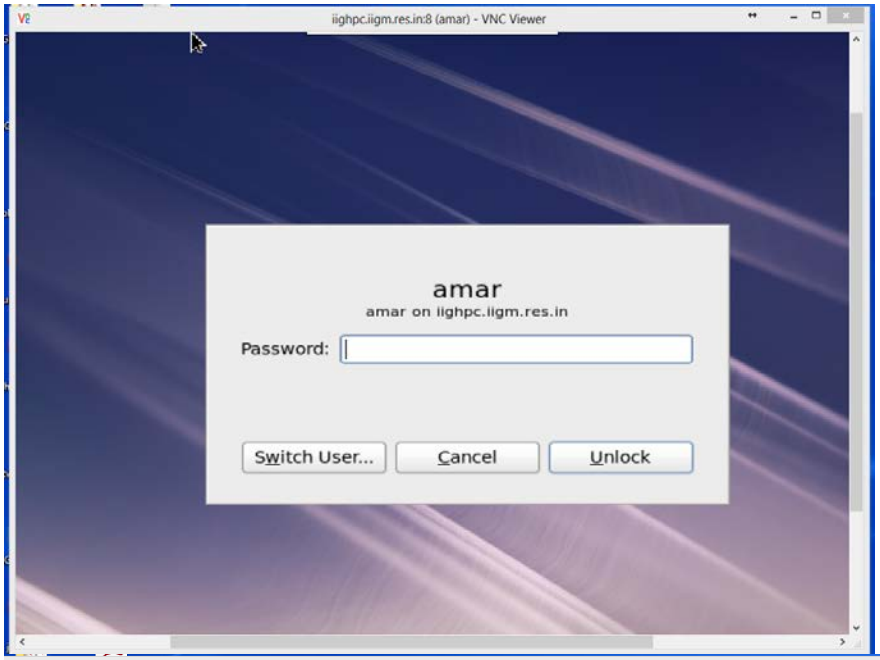


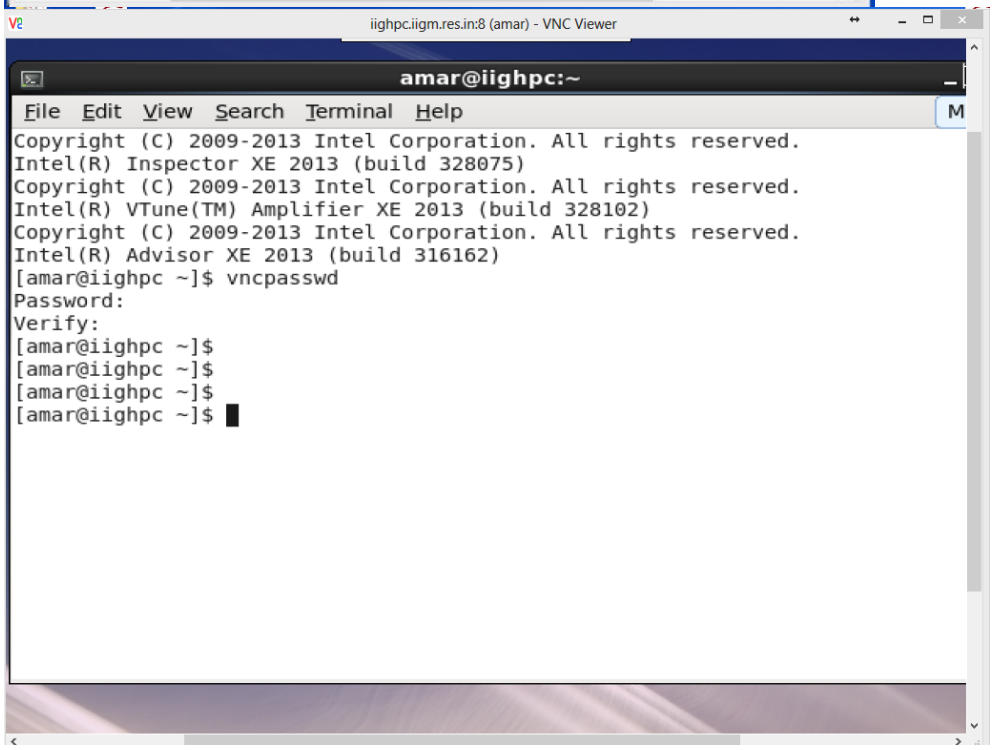


Enter your VNC Viewer password below



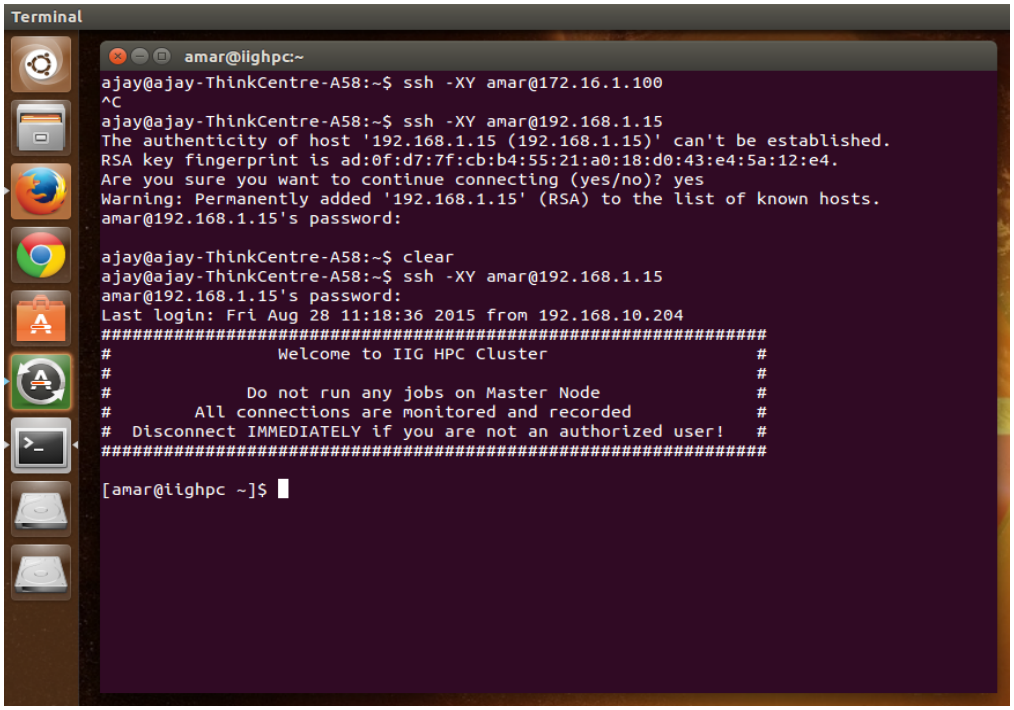
Enter your HPC login password below



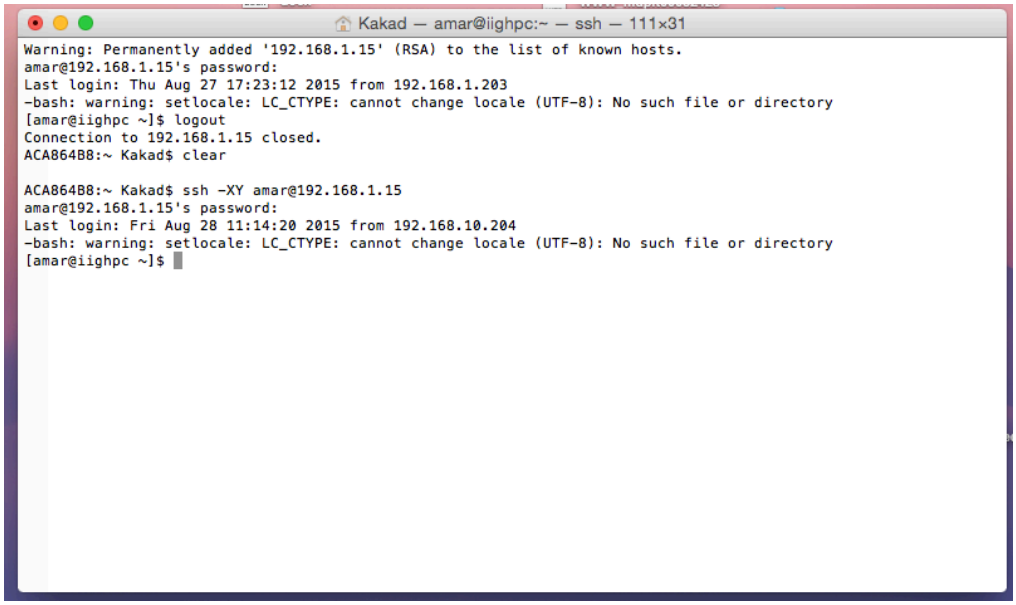


2. From Linux/Mac machine

From Linux: Open a terminal or CLI and enter, ssh –XY login-name@192.168.1.15 as shown below

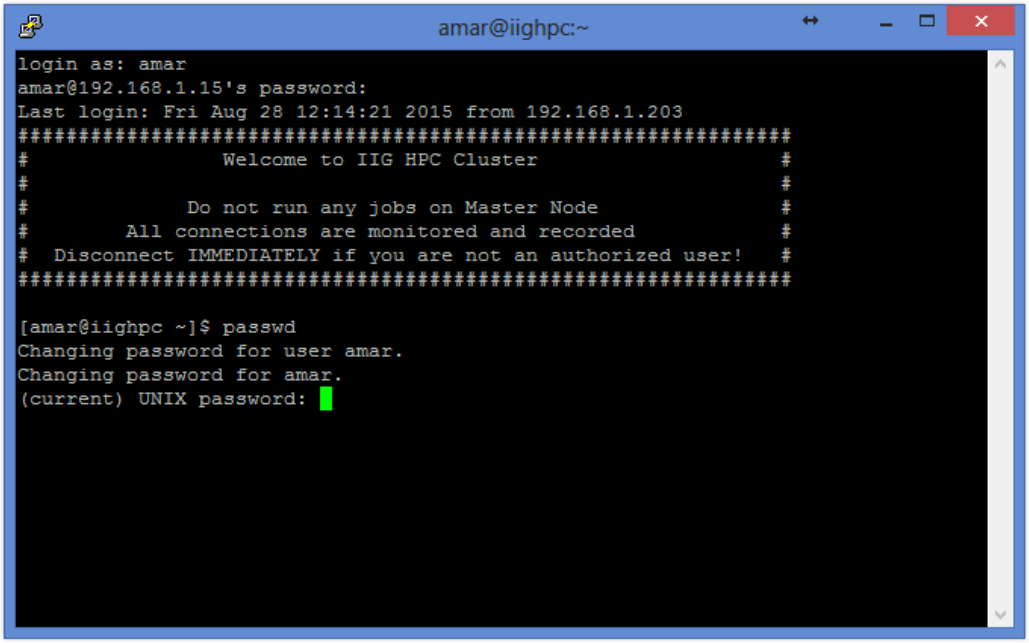


From Mac: Open a terminal or x11 and enter, ssh –XY login-name@172.16.1.100 as shown below



How to change password

Type passwd at command prompt and follow the instructions



3. How to run jobs

Basic instructions while running jobs

It is recommended to submit/run Parallel and serial jobs through UGE grid engine only i.e. using command qsub submit "script.sh". These scripts are directly submitted through master node.

Some of the examples for parallel and serial jobs are given below.

If user wish to submit a job without a script file (i.e. without qsub) then it is mandatory to submit it on a compute node only. NO JOBS SHOULD BE RUN/SUBMITTED ON THE HEAD NODE DIRECTLY.

When software like MatLab, Comsol, IDL are used it should be accessed from compute node only (node0 to node7).

You can access specific compute node of HPC cluster with following command
ssh rack1-nodeX
Here X is 0 to 7.
For e.g. to access node0, you can enter following command

ssh rack1-node0

The maximum allowed memory usage per node is 200GB and it should be strictly follow

Basic Compilers

Names	Path	Distributor
GNU gcc version 4.4.7 (gcc, g++)	/usr/bin	GNU Compiler
Intel Compilers (icc, ifort)	/opt/intel/composer_xe_2013_sp1.2.144/bin/intel64/	Intel Cluster Studio

Math Libraries

Names	Path	Description
GNU Scientific Library	/usr/lib	Mathematical libs How to use gsl on cetus?
Intel MKL Library	/opt/intel/composer_xe_2013_sp1.2.144/mkl/lib/intel64	Intel Cluster Studio Mathematical libs.

4. Application Software's on HPC

We have MATLAB, COMSOL and IDL on the HPC. Currently MATLAB and IDL can be used to run only serial jobs, whereas COMSOL is supported with parallel environment. You can access these software through only VNC viewer. Contact system administrator for the VNC viewer account.

Names	Path	Description
Matlab Version 8.1.0.604 (R2013a)	/usr/local/MATLAB/R2013a/bin	MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Matlab Homepage
IDL Version 8.2	/usr/local/exelis	IDL, short for Interactive Data Language, is a programming language used for data analysis. IDL is vectorized, numerical, and

		interactive, and is commonly used for interactive processing of large amounts of data (including image processing). The syntax includes many constructs from Fortran and some from C.
Comsol Version 5.1	/usr/local/comsol51/multiphysics/bin	COMSOL Multiphysics is a finite element analysis, solver and Simulation software / FEA Software package for various physics and engineering applications, especially coupled phenomena, or multiphysics. In addition to conventional physics-based user interfaces, COMSOL Multiphysics also allows for entering coupled systems of partial differential equations (PDEs).

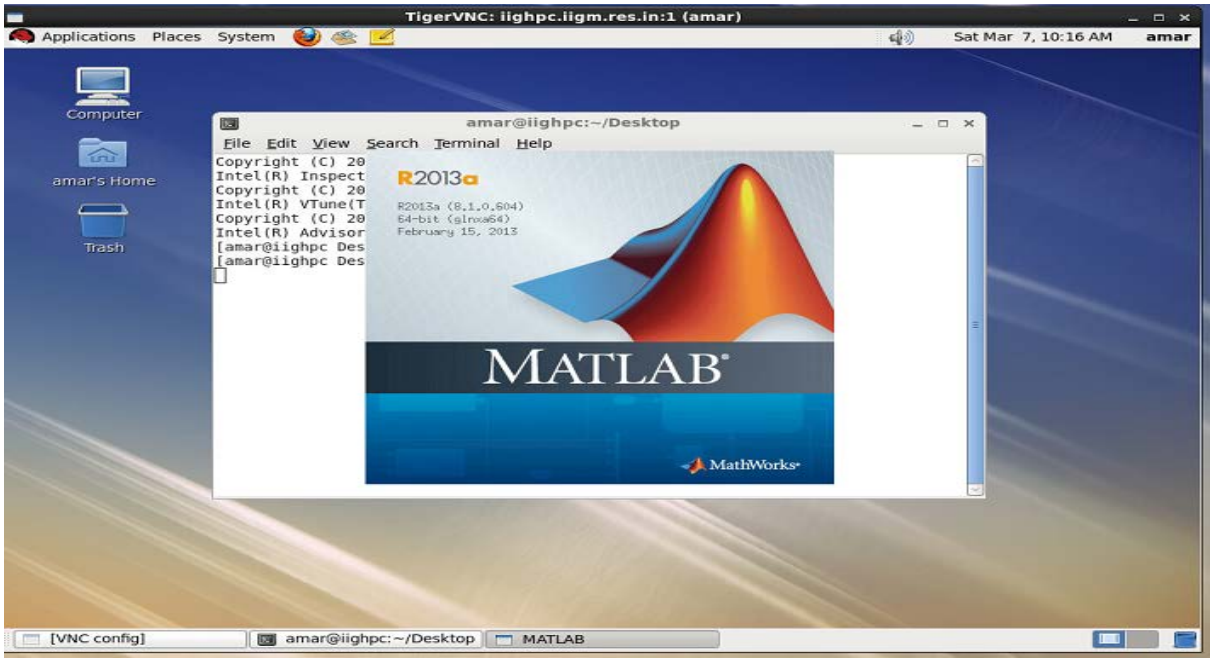
How to open Matlab:- Open VNC viewer, Click to Applications > System Tools > Terminal. In the terminal go to available compute node using command

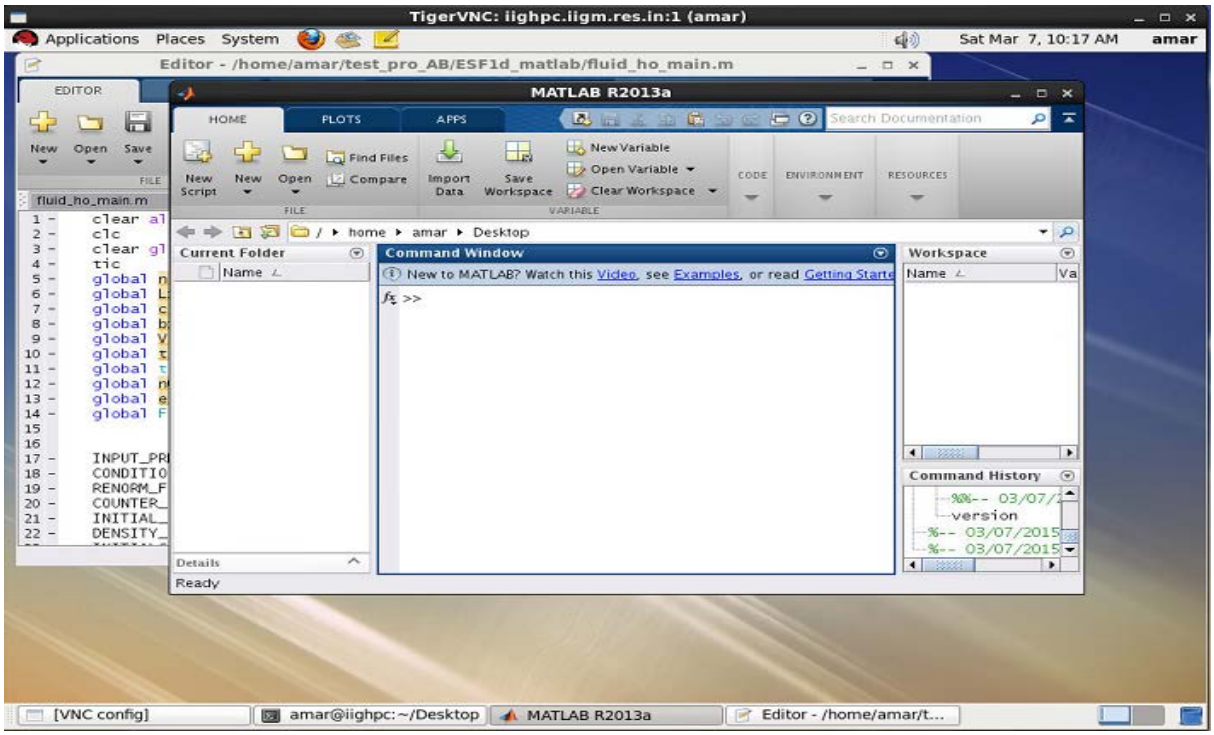
>> ssh rack1-nodeX

In the above command, X is any compute node number between 0 to7. Then type matlab at command prompt

>> matlab

Entering this command will launch the matlab GUI. Refer following snapshots for accessing matlab





How to open comsol software:

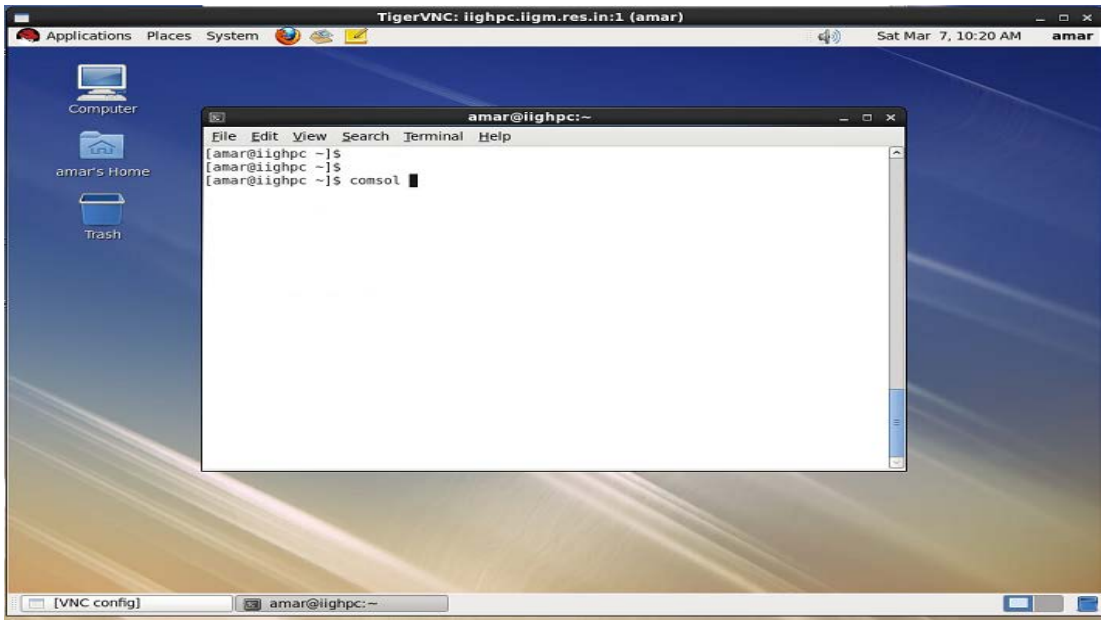
Go to any available compute node using :

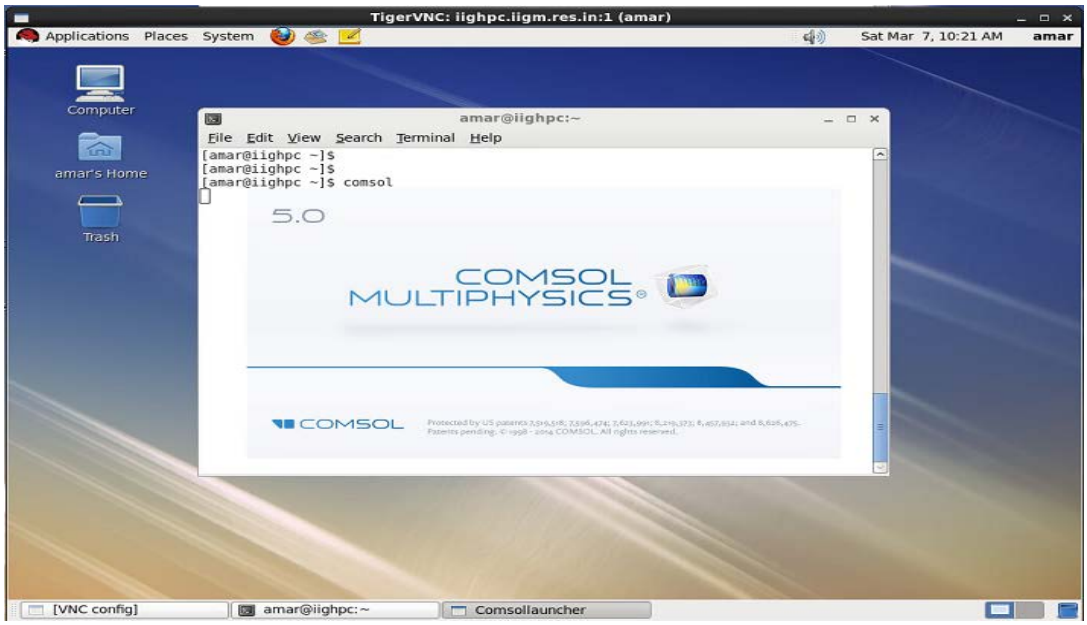
>> ssh rack1-node6

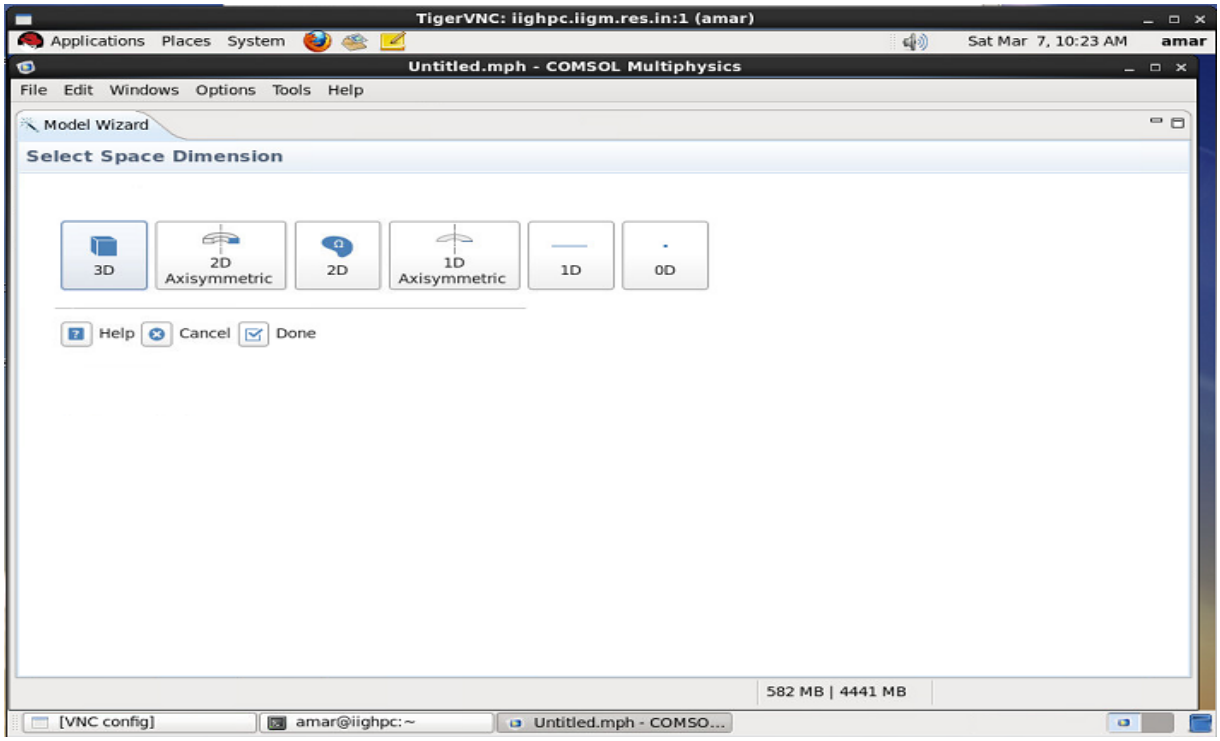
Type comsol at command prompt

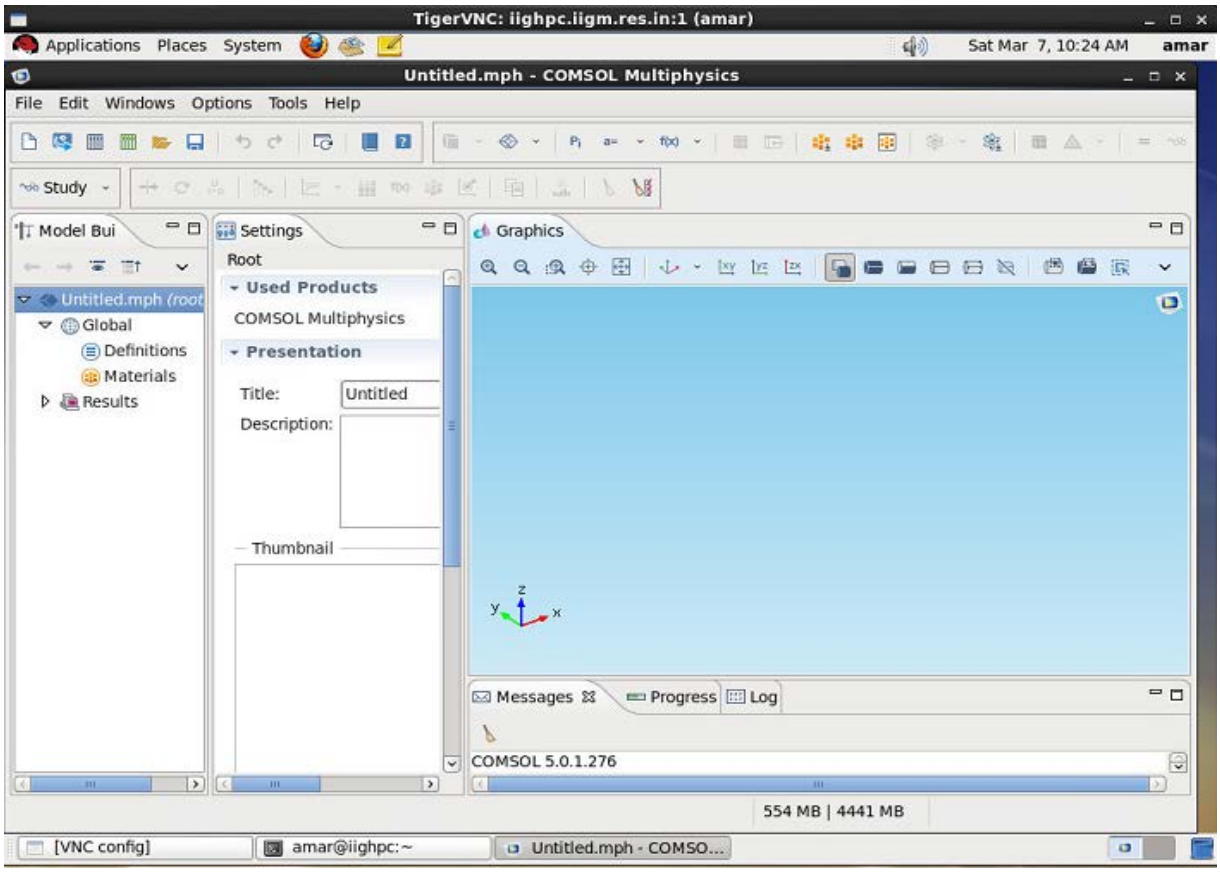
>> comsol

Entering this command will launch the the COMSOL GUI. Refer following snapshots for accessing COMSOL





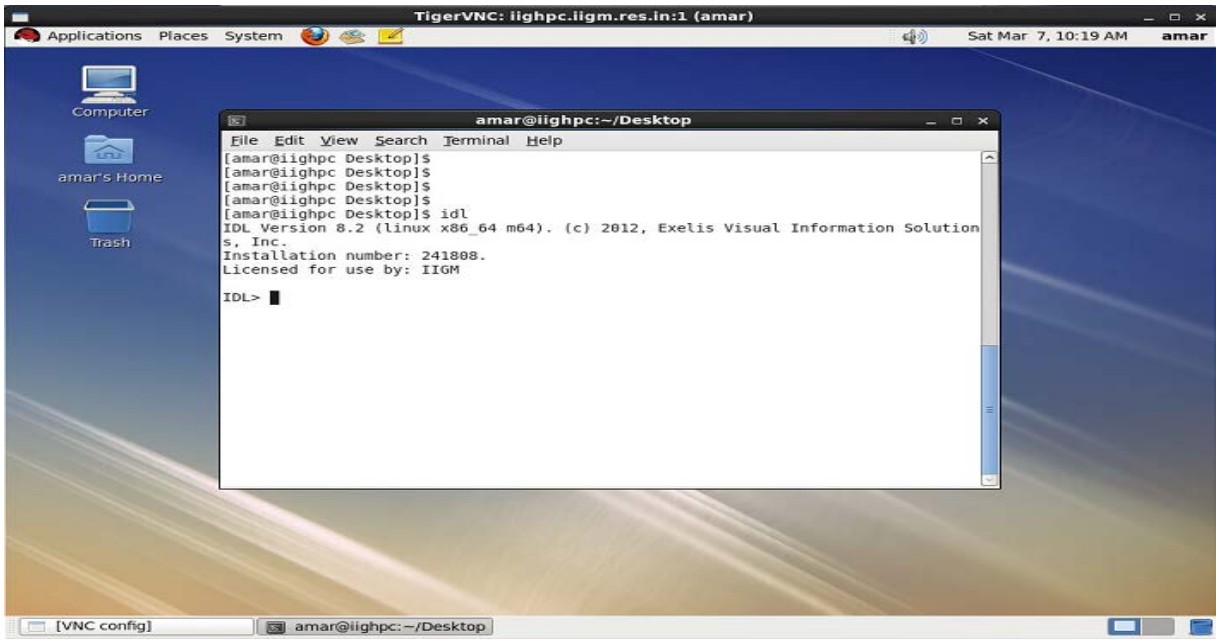

TigerVNC: iighpc.iigm.res.in:1 (amar)
Applications Places System
Sat Mar 7, 10:23 AM amar
Untitled.mph - COMSOL Multiphysics
File Edit Windows Options Tools Help
Model Wizard
Select Space Dimension
3D
2D Axisymmetric
2D
1D Axisymmetric
1D
0D
Help Cancel Done
582 MB | 4441 MB
[VNC config]
amar@iighpc:~
Untitled.mph - COMSO...


TigerVNC: iighpc.iigm.res.in:1 (amar)
Applications Places System
Sat Mar 7, 10:24 AM amar
Untitled.mph - COMSOL Multiphysics
File Edit Windows Options Tools Help
Study
Model Bui
Untitled.mph (root
Global
Definitions
Materials
Results
Settings
Root
Used Products
COMSOL Multiphysics
Presentation
Title:
Untitled
Description:
Thumbnail
Graphics
Messages
Progress
Log
COMSOL 5.0.1.276
554 MB | 4441 MB
[VNC config]
amar@iighpc:~
Untitled.mph - COMSO...

How to open IDL Software: Login to VNC viewer terminal and then go to available compute node using following command.

>> ssh rack1-node6

Then type idl at command prompt, which will launch the idl software. Presently this software works in command mode only. Here are the detail screenshot to access idl software in HPC.



5. Submitting Serial and Parallel jobs on HPC

How to submit via UGE: - (Univa Grid Engine)

A] Parallel Job Submission Script: FORTRAN code

```
#! /bin/bash
#$  -N job_name
#$  -cwd
#$  -q new.q
#$  -pe make 64
#$  -M amar@iigs.iigm.res.in
#$  -m be
# Requesting 12hrs wall clock time
#$  -l h_rt=12:00:00
#$  -l mem_free=64G
#$  -e Err.$JOB_NAME.$JOB_ID
#$  -o Out.$JOB_NAME.$JOB_ID

/opt/intel/impi/4.1.3.048/intel64/bin/mpirun  ./job_name
```

B] Serial Job Submission Script: FORTRAN

```
#!/bin/bash
#$  -N job_name
#$  -cwd
#$  -q new.q
#$  -pe make 1
#$  -M amar@iigs.iigm.res.in
#$  -m be
# Requesting 12hrs wall clock time
#$  -l h_rt=12:00:00
#$  -l mem_free=64G
#$  -o job_name.out.$job_id
#$  -e job_name.err$job_id

/opt/intel/impi/4.1.3.048/intel64/bin/mpirun ./job_name
```

In the above scripts,
-N The name you want to give the job
-pe The parallel environment of the job. This can be make, openmp
You must also specify the number of processors. This can be 2 or more
-cwd Runs the job from the directory of submission

-o File to send standard output
-e File to send standard error
-l mem_free **sets the memory requirement upper limit is 200GB**

C] Parallel Job Submission Script: COMSOL code

```
#! /bin/bash
#$  -N job_name
#$  -cwd
#$  -q new.q
#$  -pe make 64
#$  -M amar@iigs.iigm.res.in
#$  -m be
# Requesting 12hrs wall clock time
#$  -l h_rt=12:00:00
#$  -l mem_free=64G
#$  -e Err.$JOB_NAME.$JOB_ID
#$  -o Out.$JOB_NAME.$JOB_ID

export I_MPI_FALLBACK=enable

/usr/local/comsol50/multiphysics/bin/comsol batch -clustersimple -mpifabrics ofa
-inputfile /home/amar/comsol/test1/micromixer_cluster.mph -outputfile
/home/amar/comsol/test1/micromixer_cluster.$(date +%Y%m%d.%H%M%S).mph
-batchlog /home/amar/comsol/test1/micromixer.$(date +%Y%m%d.%H%M%S).log
```

Some of the basic Grid Engine Commands

How to submit a job
#qsub sge.sh

To get running job information related to cores used on nodes.
#qstat –f

To get running current job resources utilization report
 # qstat –j <job-id>

To get job information details after job done of any job-id.
 # qacct –j <job-id>

To see status of jobs.
 # qstat –u “*”
Qw: waiting in queue stat

R: running state
E: error state
T: job is transferring on cluster.

To get details job state.
 # qstat –f –u “*”

To delete the job.
 # qdel <job id>

To get list of parallel environment
qconf -spl

To get details on available parallel environment
qconf -sp openmp
qconf -sp make

To check the job submitted status & no. of cores used in cluster:-

[amar@iighpc ~]$ qstat -f

```
queuename                      qtype resv/used/tot. np_load  arch          states
---------------------------------------------------------------------------------
all.q@iighpc.iigm.res.in       BIP   0/0/12         -NA-     lx-amd64      au
---------------------------------------------------------------------------------
all.q@rack1-node0.iigm.res.in  BIP   0/0/32         0.04     lx-amd64
---------------------------------------------------------------------------------
all.q@rack1-node1.iigm.res.in  BIP   0/0/32         0.03     lx-amd64
---------------------------------------------------------------------------------
all.q@rack1-node2.iigm.res.in  BIP   0/0/32         0.04     lx-amd64
---------------------------------------------------------------------------------
all.q@rack1-node3.iigm.res.in  BIP   0/0/32         0.04     lx-amd64
---------------------------------------------------------------------------------
all.q@rack1-node4.iigm.res.in  BIP   0/0/32         0.04     lx-amd64
---------------------------------------------------------------------------------
all.q@rack1-node5.iigm.res.in  BIP   0/0/32         0.04     lx-amd64
---------------------------------------------------------------------------------
new.q@rack1-node6.iigm.res.in  BIP   0/0/32         0.04     lx-amd64
---------------------------------------------------------------------------------
new.q@rack1-node7.iigm.res.in  BIP   0/0/32         0.04     lx-amd64
[amar@iighpc ~]$
```

To check memory used each node in cluster:-

[amar@iighpc ~]$ qhost

```
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR NLOAD  MEMTOT  MEMUSE
SWAPTO  SWAPUS
----------------------------------------------------------------------------------------
global                  -               -    -    -    -     -       -       -       -
iighpc                  lx-amd64       12    2   12   12     -  62.8G       -  128.0G       -
rack1-node0             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
rack1-node1             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
rack1-node2             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
rack1-node3             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
rack1-node4             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
rack1-node5             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
rack1-node6             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
rack1-node7             lx-amd64       32    4   32   32  0.03  252.1G    1.7G  128.0G    0.0M
[amar@iighpc ~]$
```

Basic Linux Commands

#man: this command shows man page of command. #man <command>
Ex: #man ls
#ls: list information about the Files

Options
- -l: list one file per line
- -t: sort by modification time
- h: print sizes in human readable format
- -a: list hidden files

#du: estimates file space usage. Ex: #du -skh

#df : report file system disk space usage. Ex: #df -h

#top : to see the process, cpu uses, memory usage.

#ps: report a snapshot of the current processes
- -e: all processes
- -f: full

#tail: outputs the last part of files.

set the library path in /etc/ld.so.conf and then does

ldconfig: configure dynamic linker run time bindings

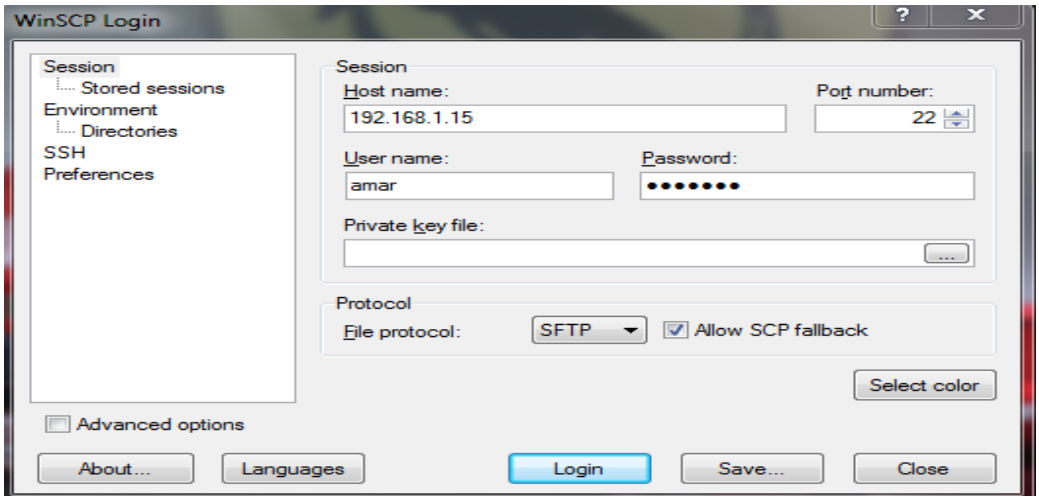
#ssh <node name>: used to login to the nodes and -XY for X forwarding. e.g. login to node 0 with X-forwarding option

>> ssh -XY usr_name@rack1-node0

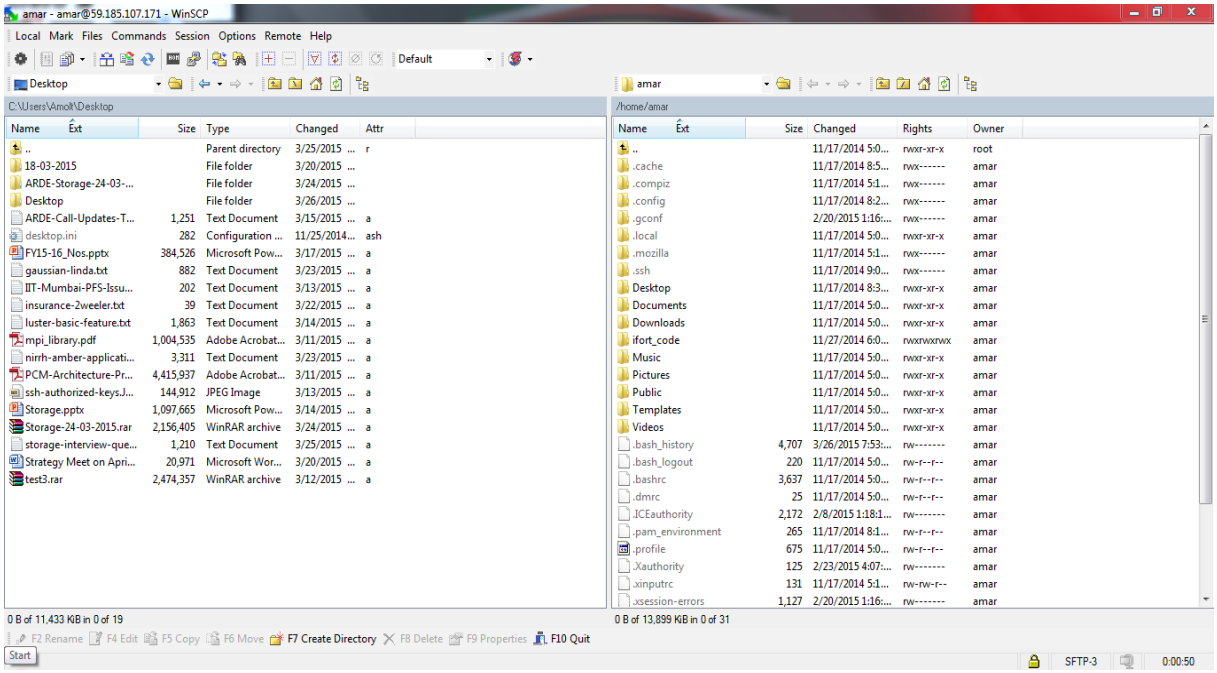
File transfer: To copy files & folder from local system to cluster user's home directory:

For Window system: User should download the WINSCP software from http://winscp.net/eng/download.php

Open the WINSCP software. Enter the hostname 192.168.1.15 with port no. 22 & username & password as shown in following screenshot



Of user, it will show your desktop contents & head node user's home directory contents



Using WINSCP, you can drag & drop, which ever folder you want to copy and paste.

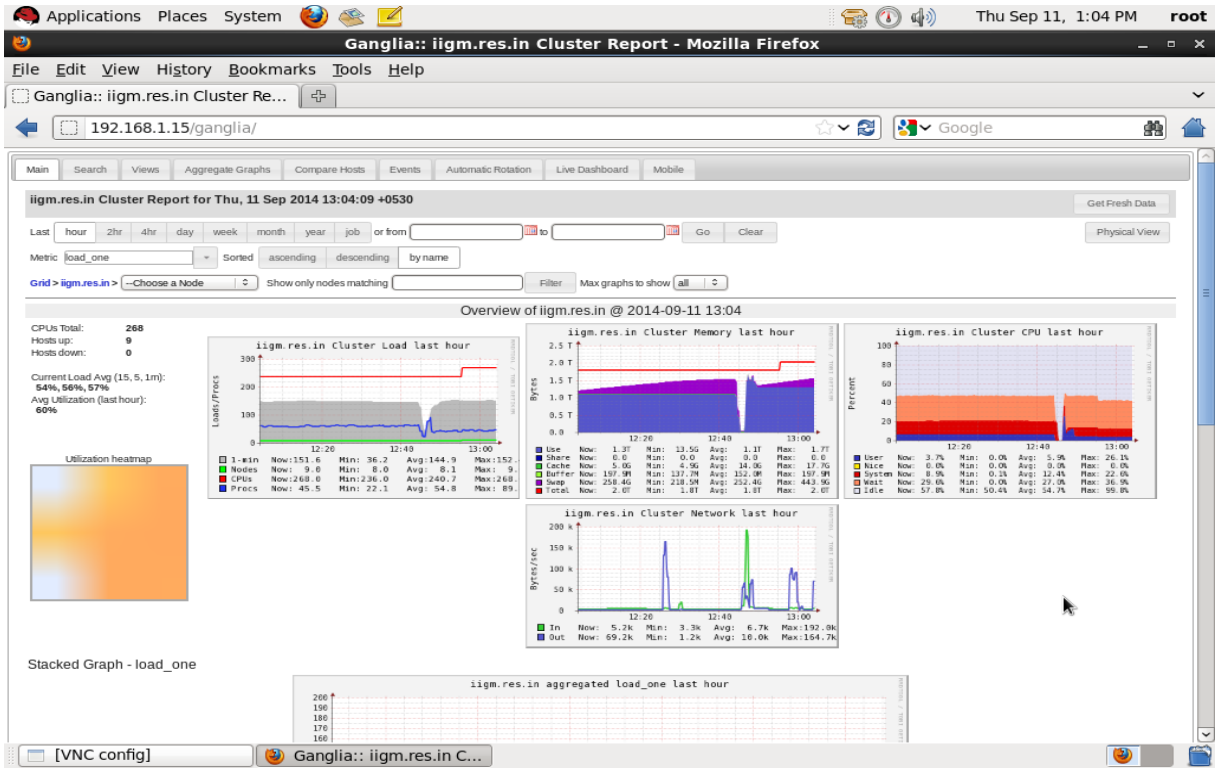
To check the storage space limit of data for the each user:-

```
[amar@iighpc ~]$ quota
Disk quotas for user amar (uid 501):
     Filesystem  blocks   quota   limit   grace   files   quota   limit   grace
172.18.1.101:/home/
                792214652       0 1073741824            7864       0       0
[amar@iighpc ~]$
```

Ganglia Cluster Monitoring Tool

A Ganglia monitoring tool provides the various information regarding the status of the Cluster. This information can be accessed through a web browser. The Ethernet IP provides the access to this monitoring tool. This monitor gathers various metrics such as CPU load, free memory, disk usage, temperature, fan speed etc. These metrics are sent through the private cluster network and are used by the master node to generate the date in a graphical manner. In addition to metric parameters, a heartbeat message from each node is collected by the Platform. When certain number of heartbeats from any node is missed, this web page will declare it "dead". A customization has been made, so as to put the temperature and Fan speed data in to this monitoring tool. The information about the Temperature and Fan speed provided by the IMM, this is accessed by the daemon ipmi (IPMI tool), which is running on all the machines. Every 30 seconds this info is fetched and given to the gmetric. The gmetric makes the information to appear in the Ganglia monitoring Page. One can launch in web browser by typing a URL.

http://localhost (from master node)
http://192.168.1.15/ganglia



Some questions and answers

1. What is the difference between make and openmp environment. When to use **make** and **openmp** environment in .sh file.
   Both are parallel environments
   make is the default parallel environment for running mpi jobs.
   openmp is the default parallel environment for running openmp jobs.

2. What is warning "Clock skew detected, your build may be incomplete". When we compile Fortran files and link those compiled files to exe.

   This warning comes due to time stamp while compilation of codes, for cleaning make build require make clean option, delete the directory containing the build, and copy fresh build code and restart the build from the beginning.

3. Is it possible to check the status of jobs given directly (without qsub)? When we close the operative (login) window, then does these direct jobs get cancelled?
   We cannot check the status of jobs submitted directly. When operative window is closed, the job will get cancelled. Hence it is recommended to submit jobs through qsub only.

4. How to give number of threads for openmp code.
   Using –pe openmp <no. of cores> in the script, but for running openmp pe, we need to have openmp compatible code.

5. How to specify particular node and cores in a .sh files
   #$ -q new.q@rack1-node6, all.q@rack1-node7 in the script
   Or
   While submitting job,
   #qsub –q all.q@rack1-node6, all.q@rack1-node7 <script name>

6. Sometimes core.****** files are generated with a very large size in the usr's account, what are these files?
   This files generates random job output. For this problem immediately contact administrator IIGHPC.

7. After completion of job, whether user is notified
   Yes, after completion of job user will get intimation. For this user need to specify his email id in the script file.

8. What is Ganglia.
   Ganglia is a scalable distributed system monitoring tool that allows remote viewing live or historical statistics for a cluster.

9. What are the available queue?

"all.q" and "new.q" are the two available queue. For high memory jobs "all.q" is used. For other jobs new.q can be used

10. How to change password of account
    Users are advised to change their password immediately after issue of a login account. Do not share passwords with anyone.

11. To submit a job on particular node (example node 6)
    qsub -q all.q@rack1-node6 script.sh

12. How to give a parallel job on command line
    mpirun -np 4 ./job_name

13. What is the limit on memory of the each node?
    200GB is the maximum limit per node. In the script file one can use #$ -l mem_free=220G.

14. How to logout
    type exit